Abstract

[0038] A method and computer program for clustering a string are described. The string includes a plurality of characters. R unique n-grams $T_{1...R}$ are identified in the string. For every unique n-gram T_S, if the frequency of T_S in a set of n-gram statistics is not greater than a first threshold, the string is associated with a cluster associated with T_S. Otherwise, for every other n-gram T_V in the string $T_{1...R, \text{ except } S}$, if the frequency of n-gram T_V is greater than the first threshold, and if the frequency of n-gram pair T_S-T_V is not greater than a second threshold, the string is associated with a cluster associated with the n-gram pair T_S-T_V. Otherwise, for every other n-gram T_X in the string $T_{1...R, \text{ except } S \text{ and } V}$, the string is associated with a cluster associated with the n-gram triple T_S-T_V-T_X. Otherwise, nothing is done.